
**PROPUESTA DE UN MODELO ESTADÍSTICO DESCRIPTIVO DE LAS VARIABLES
MÁS REPRESENTATIVAS DE LA PLATAFORMA TECNOLÓGICA DE UN SISTEMA
DE TRANSPORTE.**

Presentado por
Juan Carlos Vargas Barrera

Fundación Universitaria Los Libertadores

Facultad de Ingeniería y Ciencias Básicas

Especialización en Estadística Aplicada

Bogotá D.C, Colombia

2018

**PROPUESTA DE UN MODELO ESTADÍSTICO DESCRIPTIVO DE LAS VARIABLES
MÁS REPRESENTATIVAS DE LA PLATAFORMA TECNOLÓGICA DE UN SISTEMA
DE TRANSPORTE.**

Presentado por
Juan Carlos Vargas Barrera

En cumplimiento parcial de los requerimientos para optar al título de

Especialista en Estadística Aplicada

Asesorado por
Msc. Sebastien Lozano Forero
Profesor

Fundación Universitaria Los Libertadores
Facultad de Ingeniería y Ciencias Básicas
Especialización en Estadística Aplicada
Bogotá D.C, Colombia

2018

Nota de Aceptación

Firma del presidente del jurado

Firma del jurado

Firma del jurado

Bogotá, D.C, Enero de 2019

Las directivas de la Fundación Universitaria Los Libertadores los jurados calificadores y el cuerpo docente no son responsables por los criterios e ideas expuestas en el presente documento. Estos corresponden únicamente a los autores y a los resultados de su trabajo.

Contenido

| | |
|---|----|
| Resumen | 1 |
| Capítulo 1 Introducción..... | 3 |
| 1.1 Introducción | 3 |
| Capítulo 2 Planteamiento del problema | 5 |
| 2.1. Descripción | 5 |
| 2.2. Objetivos | 7 |
| 2.2.1. Objetivo general | 7 |
| 2.2.2. Objetivos específicos | 7 |
| 2.3. Justificación..... | 7 |
| Capítulo 3 Marco Conceptual..... | 8 |
| 3.1. Descripción. | 8 |
| 3.1.1. Operación de transporte terrestre en Colombia. | 8 |
| 3.1.2. Nuevas tecnologías aplicadas a los sistemas de transporte..... | 9 |
| 3.2. Modelos multivariados | 11 |
| Capítulo 4 Marco Metodológico..... | 12 |
| 4.1. Descripción. | 12 |
| 4.1.1. Información analizada..... | 12 |
| 4.1.2. Uso de CPU del servidor central. | 12 |
| 4.1.3. Número de conexiones al servidor central. | 12 |
| 4.1.4. Número de eventos registrados en el sistema. | 13 |
| 4.1.5. Cantidad de móviles registrados..... | 13 |
| 4.2. Unidades de las series de tiempo..... | 13 |
| 4.3. Análisis estadístico. | 14 |
| 4.3.1. Prueba de estacionalidad, test de raíces unitarias: | 14 |
| 4.3.2. Test de Ljung-Box. | 15 |
| 4.3.3. Modelos VAR (Vectores Autoregresivos). | 15 |
| Capítulo 5 Análisis y Resultados | 17 |
| 5.1. Resultados. | 17 |
| 5.2. Identificación..... | 18 |
| 5.3. Prueba de raíces unitarias..... | 18 |

| | | |
|------------|--|----|
| 5.4. | Estimación del modelo..... | 20 |
| 5.5. | Diagnóstico del modelo. | 20 |
| 5.6. | Funciones Impulso respuesta. | 25 |
| 5.7. | Descomposición de varianza. | 28 |
| 5.8. | Ajuste del modelo propuesto..... | 31 |
| 5.9. | Ajuste del modelo propuesto..... | 31 |
| 5.10. | Ecuaciones de cada una de las variables del modelo propuesto. | 32 |
| Capítulo 6 | Conclusiones y recomendaciones | 35 |
| 6.1. | Conclusiones | 35 |
| 6.2. | Recomendaciones | 35 |
| Capítulo 7 | Referencias bibliográficas | 37 |
| 7.1. | Referencias | 37 |

Índice de Tablas

| | |
|---|----|
| Tabla 1 Unidades de cada una de las variables analizadas. | 14 |
| Tabla 2: Resultados Dickey Fuller y función de auto correlación para las series. | 20 |
| Tabla 3: Diagnostico de modelos VAR. | 21 |

Índice de Imágenes

| | |
|--|----|
| Ilustración 1 Diagrama general de un Sistema de transporte. | 5 |
| Ilustración 2. Diagrama metodología Box-Jenkins. | 17 |
| Ilustración 3 Comportamiento en el tiempo de las variables seleccionadas para hacer parte del modelo | 18 |
| Ilustración 4. Pruebas raíces unitarias para la variable CPU_SC. | 19 |
| Ilustración 5. Pruebas raíces unitarias para la variable Conexiones SC. | 19 |
| Ilustración 6. Pruebas raíces unitarias para la variable Eventos. | 19 |
| Ilustración 7. Pruebas raíces unitarias para la variable Móviles. | 20 |
| Ilustración 8. Residuales para la variable CPU_SC modelo VAR(8). | 22 |
| Ilustración 9. Residuales Conexiones SC modelo VAR(8). | 22 |
| Ilustración 10. Residuales Eventos modelo VAR(8). | 23 |
| Ilustración 11. Residuales Móviles modelo VAR(8). | 23 |
| Ilustración 12. Diagnostico Ljung-Box y QQplot para los residuales de CPU_SC. | 24 |
| Ilustración 13. Diagnostico Ljung-Box y QQplot para los residuales de Conexiones SC. | 24 |
| Ilustración 14. Diagnostico Ljung-Box y QQplot para los residuales de Eventos. | 24 |
| Ilustración 15. Diagnostico Ljung-Box y QQplot para los residuales de Móviles. | 25 |
| Ilustración 16. Respuesta que presentan la CPU del sistema central cuando se presenta innovación en otra serie. | 26 |
| Ilustración 17. Respuesta que presentan a las conexiones cuando se presenta innovación en otra serie. | 27 |
| Ilustración 18. Respuesta que presentan los Eventos cuando se presenta innovación en otra serie. | 27 |
| Ilustración 19. Respuesta que presentan los Móviles en el sistema cuando se presenta innovación en otra serie. | 28 |
| Ilustración 20. Descomposición de varianza para la CPU del sistema central. | 29 |
| Ilustración 21. Descomposición de varianza para las Conexiones del sistema central. | 29 |
| Ilustración 22. Descomposición de varianza para los Eventos del sistema. | 30 |
| Ilustración 23. Descomposición de varianza para los Móviles del sistema. | 30 |
| Ilustración 24. Test Outsample para el modelo VAR(8). | 31 |
| Ilustración 25. Test outsample para cada una de las variables. | 32 |

PROPUESTA DE UN MODELO DESCRIPTIVO DE LAS VARIABLES MÁS REPRESENTATIVAS DE LA PLATAFORMA TECNOLÓGICA DE UN SISTEMA DE TRANSPORTE

Resumen

Los sistemas de transporte actuales incluyendo público y de carga necesitan para su operación una plataforma tecnológica que reciba, procese y analice todo tipo de información enviada por los equipos a bordo con el fin de controlar y optimizar la operación de los móviles que la componen. Estas plataformas tecnológicas tienen variables en común que son representativas en todos los sistemas de transporte, debido a su importancia estas variables son monitoreadas por el equipo técnico para detectar comportamientos anormales y al observarlas en tiempo real permiten detectar posibles fallos en algún componente de la plataforma. Este procedimiento manual de verificación no cuenta con un respaldo teórico, se ha ido alimentando de la experiencia de los ingenieros de soporte y de las fallas ocurridas anteriormente.

Con el fin de brindar un respaldo teórico desde el punto de vista estadístico se propone modelar el sistema a través de un vector autoregresivo VAR(8) utilizando 4 de las variables más representativas del sistema: uso de CPU, cantidad de conexiones, cantidad de eventos, cantidad de móviles en operación.

Como resultado se obtuvo información valiosa acerca de cada una de las variables y como entre las 4 variables escogidas se destacan las conexiones al sistema y la cantidad de eventos, esta información nos permitió mejorar y enfocar nuestras herramientas de monitoreo sobre estas dos variables sin desperdiciar tiempo ni recursos en las demás.

Palabras clave: Series de tiempo, sistema de transporte, modelo VAR, modelo predictivo.

Capítulo 1 Introducción

1.1 Introducción

Los sistemas de transporte terrestre son los principales pilares en la movilidad de ciudadanos, bienes y servicios en las grandes ciudades, estos sistemas articulan cada una de las actividades realizadas por los sectores productivos, por esta razón se consideran a la vez uno de los factores más importantes para la productividad y el desarrollo ya que a medida que se hacen más eficientes los sistemas de transporte terrestre se incrementa a su vez la eficiencia de los sectores productivos y de prestación de servicios. Es así que la movilidad se convierte en uno de los ítems destacados para la equidad y democratización de los derechos universales de Libre Circulación, seguridad, igualdad de oportunidades, de acceso a bienes y servicios, en conclusión para mejorar la calidad de vida de la población en su conjunto.

La Ingeniería de Sistemas, en los últimos años ha brindado gran parte de sus esfuerzos al manejo y análisis de la información en el sector de transporte, su objetivo es dotar a este tipo de sistemas de inteligencia que le permita adaptarse, optimizarse y aprender del entorno es por esto que vemos grandes desarrollos de técnicas de planificación de rutas y administración de recursos para hacer a los sistemas de transporte más eficientes. Esta disciplina, además, ha hecho posible la integración de los modos de transporte, mediante su interacción ordenada (transportación intermodal y multimodal), lo cual evita competencias inútiles y un mejor aprovechamiento de las características de cada uno de ellos (Islas Rivera & Zaragoza, 2007), por otra parte a través de los nuevos desarrollos en software se han aplicado nuevas técnicas para el procesamiento y almacenamiento de los datos de estos sistemas, permitiendo a su vez obtener una gran cantidad de información en tiempo real la cual es usada para gestión de los móviles y alimentar su aprendizaje.

Todo este potencial de software se respalda bajo equipos físicos, entre los cuales encontramos, servidores, almacenamiento y enrutadores todos ellos de alto rendimiento que permiten realizar cada una de las tareas de la aplicación en tiempos mínimos con un alto grado de confiabilidad.

Teniendo en cuenta estos antecedentes, el objetivo del presente trabajo consiste en proponer un modelo de pronóstico para cada una de las variables de la plataforma tecnológica de un sistema de transporte que son monitoreadas. Para ello se cuenta con cuatro conjuntos de datos suministrados por las herramientas de monitoreo que actualmente se encargan de analizar los parámetros considerados primordiales durante la operación, toda la información es guardada en diferentes plataformas como bases de datos, archivos planos y logs.

Analizando las series se obtuvo un modelo VAR que modela con precisión cada una de las variables monitoreadas y con el que además se pronosticó cada una de ellas, encontrando que su pronóstico es igualmente acertado, luego de este análisis el departamento de Ingeniería fijará los límites operativos de estas variables y de acuerdo a estos se determinará si se considera una alerta.

Esta investigación se divide en cuatro partes, la primera es el marco conceptual donde se dará a conocer los temas que han sido desarrollados relacionados al eje central de esta investigación. En segundo lugar se describirá la metodología con la cual se obtuvieron los datos y los métodos estadísticos usados para realizar el análisis, seguido a esto se presentaran los resultados obtenidos y el análisis de los mismos, por último se presentaran las conclusiones de la investigación, las recomendaciones para futuros investigadores y las referencias bibliográficas.

Capítulo 2 Planteamiento del problema

2.1. Descripción

En la industria del transporte las plataformas tecnológicas cumplen un papel fundamental en la gestión de los recursos, ya que sobre estas operan cada una de los componentes que conforman el sistema y de la cual depende de manera completa la gestión y monitoreo de la flota, por esta razón la plataforma tecnológica debe ser confiable, segura y permanecer disponible durante todo el horario de operación.



Ilustración 1 Diagrama general de un Sistema de transporte.

Actualmente, el manejo y análisis de la información se ha vuelto indispensable para todas las empresas, el desarrollo de las nuevas tecnologías ha permitido la interacción de diferentes ciencias en campos distintos a los que comúnmente eran aplicables, esto ha permitido que la visión dada por los resultados del análisis se amplíe y brinde información general pero detallada acerca de un interés específico. Es así que en esta investigación se combinan la extracción y depuración de datos con herramientas de ingeniería y el análisis de la información dado por la estadística.

Aunque las plataformas tecnológicas están siendo monitoreadas las 24 horas y los 7 días de la semana durante todo el año, al estar compuestas de equipos físicos y manipulada por humanos no es exenta de algún tipo de fallo intrínseco a cada componente, pueden ser programaciones de ruta

que al no ser generadas de manera correcta generan ciclos infinitos que hacen que el software bloquee el sistema central, igualmente los servidores de sistema central o de base de datos pueden verse afectados por tareas distintas a las relacionadas directamente con la operación que incrementan el uso de su procesamiento lo que se refleja luego en lentitud del sistema y/o en la inserción de datos.

Cada falla presentada en la plataforma tecnológica repercute en situaciones que afectan las empresas como por ejemplo, móviles trabajando sin control, pérdida de información de recorridos realizados, pérdida de integridad de la información para reportes, disminuye la disponibilidad del sistema, la calidad del servicio se ve afectada, muchas veces se generan pleitos por la confiabilidad de la información generada durante las fallas, entre otras, cada una de estas fallas generan gasto económicos y de personal que las empresas relacionadas deben asumir.

Luego de observar esta situación en diferentes empresas de transporte, se planteó la siguiente pregunta de investigación: ¿Qué modelo estadístico describe el comportamiento de las variables más representativas de la plataforma tecnológica de un Sistema de transporte?

2.2. Objetivos

2.2.1. Objetivo general

Modelar el comportamiento de las variables representativas de un sistema de transporte de pasajeros o carga través un modelo estadístico.

2.2.2. Objetivos específicos

- Determinar las variables más representativas en la operación de un sistema de transporte.
- Comparar diferentes modelos predictivos y modelar las variables con el mejor obtenido.
- Predecir los 24 valores futuros de la plataforma para cada una de las variables.

2.3. Justificación

Una empresa de transporte tiene la necesidad de conocer el comportamiento futuro de sus variables más representativas con el fin de prever fallas que afecten de manera permanente el sistema e influyan de manera negativa la gestión y afecten la información generada y almacenada en la plataforma, este tipo de comportamiento terminará afectando de alguna manera las finanzas de la compañía. Para pronosticar con precisión el comportamiento de las variables se deben conocer los valores usuales de cada variable durante la operación y de acuerdo a estos valores determinar límites que nos informen acerca de comportamientos inusuales en la plataforma tecnológica.

Capítulo 3 Marco Conceptual

3.1. Descripción.

3.1.1. Operación de transporte terrestre en Colombia.

La operación del transporte de carga para Colombia se evalúa mediante el cálculo de una serie de indicadores, que permiten medir la forma en que está operando la movilización de mercancías en el territorio nacional.

Para la elaboración de estos indicadores, se relacionan una serie de datos referentes al parque automotor, la movilización de carga y la demanda potencial de transporte.

- Oferta Vehicular: Está constituida por toda la información referente al parque automotor de transporte de carga en Colombia, cuantificada en la cantidad de vehículos y por el número de toneladas que se pueden movilizar en estos vehículos.
- Demanda Potencial de Transporte: Se considera la demanda como la cantidad de productos, medida en toneladas, que están en potencia de ser transportados en el territorio nacional.

Movilización de Carga: El análisis muestra la manera en que se está movilizand la carga por las carreteras del país, mediante preguntas directas al conductor e inspección visual del vehículo. Básicamente se evalúa la cantidad de carga movilizad, los tipos de vehículos empleados y la cantidad de viajes realizados entre otros.

- Sistemas de Información: La consolidación de información para evaluar la Operación del Transporte de Carga por Carretera, se realiza mediante un sistema de manejo de datos denominado “SISTEMA DE INFORMACIÓN DE OPERACIÓN DEL TRANSPORTE DE CARGA - SIOTCA”. Está compuesto por tres modelos generadores de información básica que son:

- Oferta vehicular.
- Demanda de transporte.
- Movilización de Carga.

Los cuales se utilizan como insumos de análisis y la evaluación se realiza mediante la formulación de indicadores de operación, que resultan de interactuar las variables obtenidas de los módulos de oferta, demanda y movilización. La salida general del sistema consiste en una serie de indicadores que son los datos, a través de los cuales se efectúa el análisis de la Operación del Transporte de Carga por Carretera en Colombia. (Canal Mora, 2001).

3.1.2. Nuevas tecnologías aplicadas a los sistemas de transporte.

La tecnología big data en la industria del transporte por carretera está volviéndose parte primordial para su gestión y control, es por eso que día a día se está haciendo evidente una brecha cada vez mayor entre la formulación de políticas y los objetivos que estas deben alcanzar. Lo que está conduciendo a resultados no deseados y está creando desafíos para las autoridades encargadas de hacer cumplir la reglamentación en la totalidad de los camiones que circulan. La discusión actual sobre la gran cantidad de datos en el transporte y su constante avance hacia los enfoques de gobernanza digital pueden presentar una oportunidad para intentar disminuir esta brecha. Esto se puede lograr a través de normativas más específicas, flexibles y mecanismos de aplicación de estas normas basadas en indicadores de actividad y rendimiento que a su vez sean cuantificables. Otro importante desarrollo relacionado es la automatización de la operación de los vehículos de carga. Esto afectará la forma en que se organizarán estos vehículos como a la industria de la logística, y será necesario regularlos de algún modo, de igual se debe brindarles un entorno con altos niveles de conectividad a redes no cableada, de esta manera, podrán generar y transmitir datos que brinden

todas las características necesarias para el éxito de la gobernanza de manera digital. Estos nuevos desarrollos nos llevarán hacia un enfoque de formulación de políticas basado en datos. Para esta actividad sea cuantificable y que los indicadores de rendimiento sean fiables tendrían que definirse junto con el acceso a fuentes de datos comerciales o privadas combinadas con contenidos de datos que ya están en el dominio público e igualmente otros datos. Dado el uso de estos datos para fines regulatorios, se requiere un alto nivel de confiabilidad y consistencia para que la información pueda ser correctamente procesada. Por lo tanto tener un sistema big data amplio pero menos estructurado podría ser un inconveniente para su implementación.

Los desarrollos actuales en el campo de la ciencia de los datos que lo componen diferentes áreas como machine learning, minería, inteligencia artificial, toma de decisiones, entre otros, podrían ser necesarios como base para implementar estas tecnologías con éxito. Por esta razón la tecnología blockchain, que actualmente recibe mucha atención en varios sectores, está tomando protagonismo. Blockchain combina avances en ciencia de datos, criptografía y nuevos principios de gerencia. Específicamente, permite transacciones confidenciales con respecto al pago y acceso a servicios, datos y derechos distribuidos sin la necesidad de que un tercero establezca un vínculo pero garantizando la seguridad de los datos entre las partes que realizan la transacción. Por otro lado abre nuevas posibilidades para administrar sin problemas la distribución tanto para vehículos como para los nodos de la infraestructura, y ofrece la posibilidad de un transporte personalizado, dinámico y de tamaño adecuado para particulares. Es así que las nuevas aplicaciones del sector transporte pueden proporcionar una solución a muchas de las preocupaciones actuales en esta área, pero también se deben abordar otros problemas principalmente son los más importantes la protección de la privacidad y el delito cibernético.

El desafío es diseñar reglas generales de modo que no discrimine entre los componentes en función de las características tecnológicas de sus flotas de vehículos. Aquí también juega un papel principal el tiempo de implementación de las tecnologías y el creciente desfase temporal de las políticas acerca del uso de datos y las normativas a estos desarrollos. Otros desarrollos relacionados, que afectarán tanto la necesidad como la oportunidad de gobernanza digital del transporte de mercancías por carretera, incluyen plataformas de análisis de demanda y el uso de vehículos más grandes y pesados (Ideas de transporte y movilidad para América Latina y el caribe, 2018)

3.2. Modelos multivariados

Los objetivos de un modelo de regresión puede ser dos:

- Obtener una ecuación que nos permita “predecir” el valor de Y una vez conocidos los valores de $X_1, X_2 \dots X_k$. Se conocen como modelos predictivos.
- Cuantificar la relación entre $X_1, X_2 \dots X_k$ y la variable Y con el fin de conocer o explicar mejor los mecanismos de esa relación. Se trata de modelos explicativos, muy utilizados cuando se busca encontrar qué variables afectan a los valores de un parámetro específico.

Capítulo 4 Marco Metodológico.

4.1. Descripción.

4.1.1. Información analizada.

La plataforma de un sistema de transporte posee varios parámetros que están siendo monitoreados por diferentes herramientas de software y/o equipos, algunos de estos parámetros son: Uso de CPU y memoria del servidor del sistema central, ancho de banda de los canales de comunicación, número de operadores registrados en la plataforma, número de móviles registrados en la plataforma, número de eventos registrados en la base de datos, uso de CPU y memoria del servidor de base de datos, uso de CPU y memoria de servidor de puerta de enlace, etc. Sin embargo de todos estos parámetros los más relevantes técnicamente, ya que cuando ocurre una falla estos parámetros se alteran de manera considerable, y a los cuales se limitó esta investigación son los cuatro siguientes:

4.1.2. Uso de CPU del servidor central.

El uso de CPU es una medida sobre qué cantidad del procesador general se utiliza en el servidor en un momento dado. Los procesos de las aplicaciones se añaden a una cola y después se alimentan a los múltiples núcleos de la CPU para ser procesados en paralelo. Cada núcleo procesa de forma independiente los datos que se le envían, los resultados después se promedian con el resto de núcleos de procesador y la salida general se puntúa desde 0% a 100%. En caso de falla el uso de CPU se incrementará de manera exponencial o disminuirá de la misma manera.

4.1.3. Número de conexiones al servidor central.

Este es el número de conexiones que establece el sistema central con cada uno de los componentes de la plataforma como por ejemplo los móviles (buses, motos, camiones, etc), los operadores del

centro de control, las aplicaciones y los demás servidores y equipos de la plataforma. En caso de falla las conexiones al sistema central disminuirán.

4.1.4. Número de eventos registrados en el sistema.

Durante la operación del sistema se presentan diferentes eventos, cada empresa los fija de acuerdo a su operación e intereses, algunos de estos eventos son: Emergencia, desvío de ruta, falla mecánica, accidente, inicio de ruta, fin de ruta, etc. En caso de falla los eventos no pueden llegar al sistema central y esta cantidad disminuye o se hace nula.

4.1.5. Cantidad de móviles registrados.

Esta variable nos indica la cantidad de móviles (vehículos automotores) que se encuentran registrados en el sistema, en caso de falla, los móviles no pueden registrarse y esta cantidad disminuye.

Por las razones descritas se escogieron estas variables como representativas y generales de un sistema de transporte.

Los datos objeto de estudio fueron recolectados de las herramientas de monitoreo instaladas en la plataforma para recolectar la información de cada uno de los parámetros. Se tomaron valores promedios para los 5 días laborales (lunes a viernes) para los últimos 4 meses (Enero a Abril 2018).

4.2. Unidades de las series de tiempo.

Durante 5 días tomamos muestras cada hora de cada una de las variables, es decir que para cada variable contamos con 120 datos que equivalen a las horas totales de los 5 días. A continuación se describen las unidades de medida de cada una de las variables:

Tabla 1 Unidades de cada una de las variables analizadas.

| VARIABLE | UNIDAD DE MEDIDA |
|-----------------------------------|-------------------------|
| Uso de CPU del servidor Central | Porcentaje de uso (%) |
| Cantidad de Conexiones al Sistema | Unidades |
| Cantidad de eventos | Unidades |
| Cantidad de móviles | Unidades |

4.3. Análisis estadístico.

Los datos obtenidos fueron sujetos a un análisis estadístico, haciendo uso del software R versión 3.4. Una vez ingresados los datos en el software se llevaron a cabo los siguientes análisis, usando un nivel de significancia del 5 % para rechazar la hipótesis nula en todas las pruebas:

4.3.1. Prueba de estacionalidad, test de raíces unitarias:

Cuando una serie no es estacionaria en media, o lo que es lo mismo, cuando no es integrada de orden cero I (0) se dice que presenta al menos una raíz unitaria. Cuando esto ocurre, sabemos que es posible la obtención de una serie estacionaria mediante una sencilla transformación de la serie original, como es la diferenciación adecuada. Entonces, el número de diferencias que habrá que tomar en la serie para convertirla en estacionaria en media viene dado, justamente, por el número de raíces unitarias que la serie original presente. Pues bien, el número de raíces unitarias de la serie se determina formalmente mediante los procedimientos de Dickey Fuller.

4.3.2. Test de Ljung-Box.

Se utilizó el estadístico q de Ljung-Box para comprobar si las series de observaciones en un período de tiempo específico son aleatorias e independientes. Si las observaciones no son independientes, una observación puede estar correlacionada con otra observación k unidades de tiempo después, una relación que se denomina autocorrelación. Esta puede reducir la exactitud de un modelo predictivo basado en el tiempo, como la gráfica de series de tiempo, y conducir a una interpretación errónea de los datos.

4.3.3. Modelos VAR (Vectores Autoregresivos).

Utilizamos un modelo del tipo vector auto regresivo (VAR) cuando queremos caracterizar las interacciones simultáneas entre un grupo de variables. Un VAR es un modelo de ecuaciones simultáneas formado por un sistema de ecuaciones de forma reducida sin restringir. Que sean ecuaciones de forma reducida quiere decir que los valores contemporáneos de las variables del modelo no aparecen como variables explicativas en ninguna de las ecuaciones. Por el contrario, el conjunto de variables explicativas de cada ecuación está constituido por un bloque de retardos de cada una de las variables del modelo. Que sean ecuaciones no restringidas significa que aparece en cada una de ellas el mismo grupo de variables explicativas. Pueden incluirse también como variables explicativas algunas variables de naturaleza determinista, como una posible tendencia temporal, variables ficticias estacionales, o una variable ficticia de tipo impulso o escalón, que sirve para llevar a cabo un análisis de intervención en el sistema. Por último, podría incluirse como explicativa una variable, incluso en valor contemporáneo, que pueda considerarse exógena respecto a las variables que integran el modelo VAR.

El modelo VAR es muy útil cuando existe evidencia de simultaneidad entre un grupo de variables, y que sus relaciones se transmiten a lo largo de un determinado número de períodos. Al no imponer ninguna restricción sobre la versión estructural del modelo, no se incurre en los errores de especificación que dichas restricciones pudieran causar al ejercicio empírico. De hecho, la principal motivación detrás de los modelos VAR es la dificultad en identificar variables como exógenas, como es preciso hacer para identificar un modelo de ecuaciones simultáneas.

En el modelo VAR, cada variable se expresa como una función lineal de sus propios valores pasados, de todas las demás variables y de un término de error estocástico.

Capítulo 5 Análisis y Resultados

5.1. Resultados.

A continuación se muestra el análisis hecho a cada una de las series. Se usó la metodología Box – Jenkins la cual consiste en 4 etapas generales que deben ser desarrolladas de manera consecutiva:

1. Identificación.
2. Estimación.
3. Verificación o diagnóstico.
4. Pronóstico.

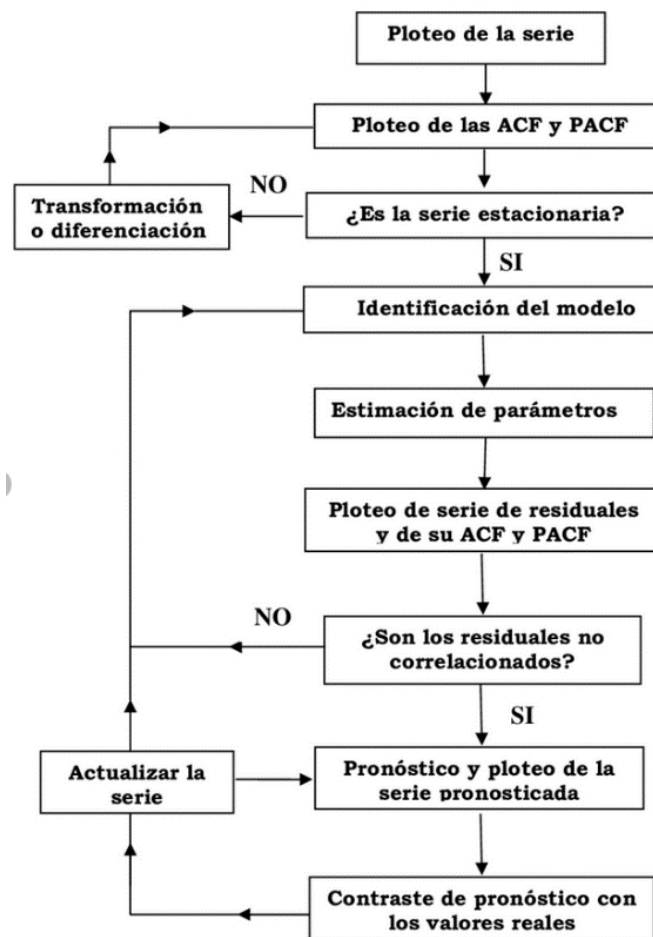


Ilustración 2. Diagrama metodología Box-Jenkins.

5.2. Identificación.

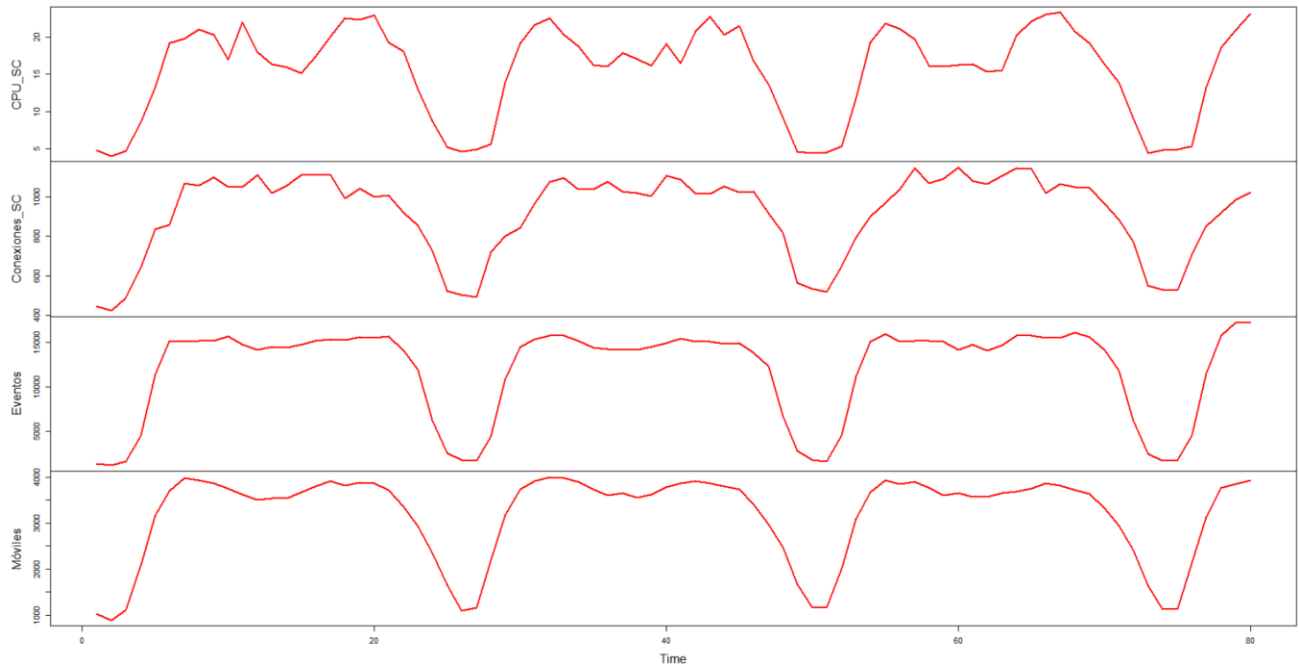


Ilustración 3 Comportamiento en el tiempo de las variables seleccionadas para hacer parte del modelo

5.3. Prueba de raíces unitarias.

Inicialmente se realiza este test para comprobar que cada una de las variables siguen una distribución normal independiente del tiempo, es decir que su varianza, media y tendencia son constantes en el tiempo.

Teniendo un valor de significancia 0.05 y la hipótesis nula e hipótesis alternativa como:

H_0 = La serie es no estacionaria, tiene raíz unitaria.

H_a = La serie es estacionaria, no tiene raíz unitaria.

Se realiza el test de dickey fuller para cada una de las series y se observa las funciones de auto correlación para cada una de ellas:

- CPU_SC (CPU del sistema central).

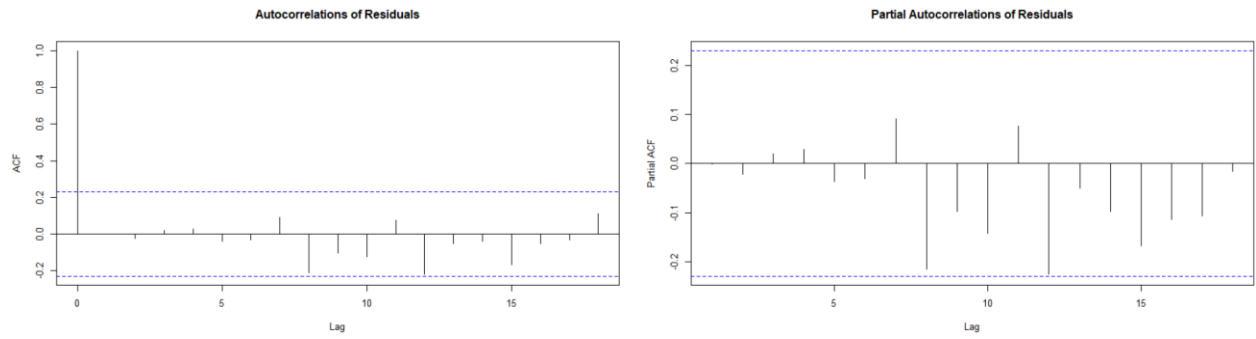


Ilustración 4. Pruebas raíces unitarias para la variable CPU_SC.

- Conexiones SC (Conexiones al sistema central).

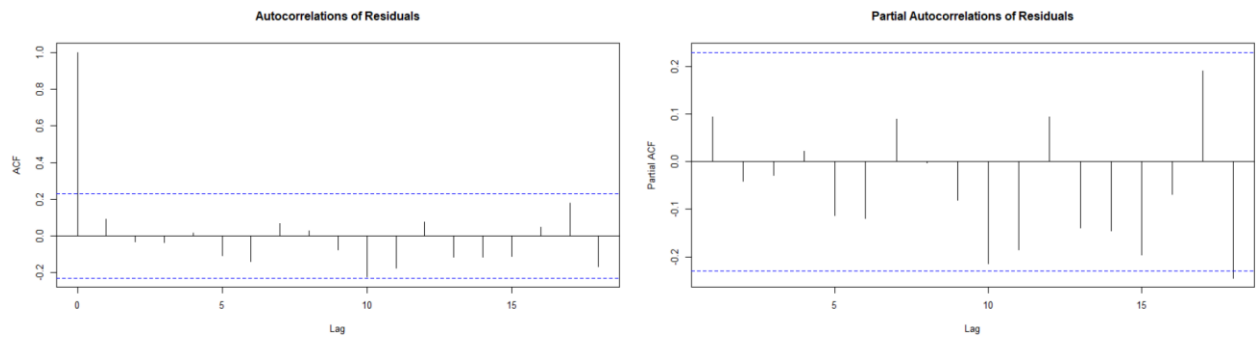


Ilustración 5. Pruebas raíces unitarias para la variable Conexiones SC.

- Eventos.

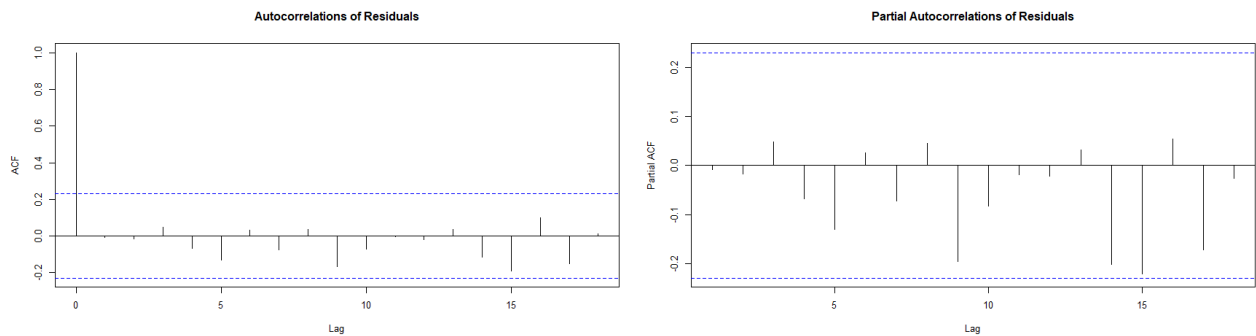


Ilustración 6. Pruebas raíces unitarias para la variable Eventos.

- Móviles.

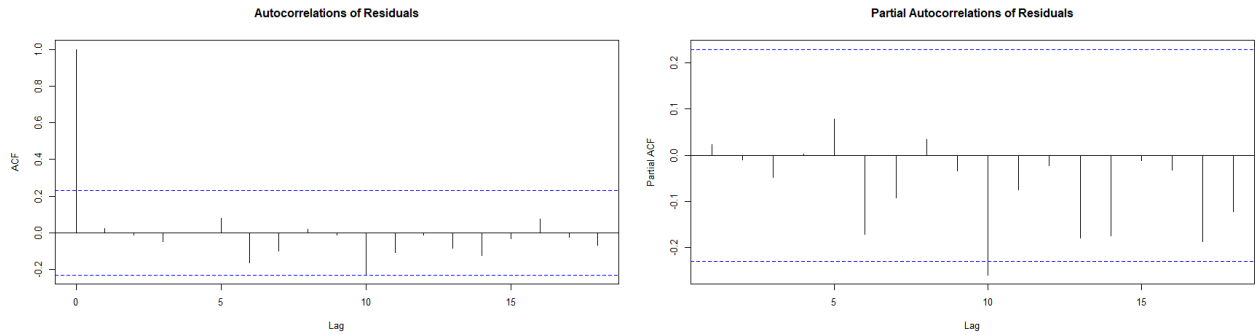


Ilustración 7. Pruebas raíces unitarias para la variable Móviles.

Se observa que cada una de las variables no existe autocorrelación de los residuales. Lo que nos garantiza que nuestras pruebas serán confiables.

Tabla 2: Resultados Dickey Fuller y función de auto correlación para las series.

| Serie | Dicky-Fuller p-value | ACF Residuales |
|---------------|------------------------|----------------|
| CPU_SC | $5.391 \cdot 10^{-8}$ | Entre bandas |
| Conexiones_SC | 0.0002102 | Entre bandas |
| Eventos | $1.481 \cdot 10^{-14}$ | Entre bandas |
| Móviles | $2.2 \cdot 10^{-16}$ | Entre bandas |

Al realizar la prueba de dicke-fuller al 5% se rechaza la hipótesis nula y se acepta la hipótesis alternativa, es decir que cada una de las series son estacionarias.

5.4. Estimación del modelo.

Inicialmente se plantea un modelo VAR (8) con un parámetro estacional de 24.

5.5. Diagnóstico del modelo.

Se realizó el diagnóstico del parámetro BIC, autocorrelacion y normalidad a 8 modelos VAR.

Tabla 3: Diagnostico de modelos VAR.

| Modelo | BIC | Serial | Normalidad JB |
|---------|----------|------------------------|---------------|
| VAR(1) | 3354.238 | 0.007992 | 0.02368 |
| VAR(2) | 3355.112 | 0.002044 | 0.05899 |
| VAR(3) | 3343.705 | 0.02637 | 0.598 |
| VAR(4) | 3312.153 | 0.006579 | 0.1472 |
| VAR(5) | 3277.285 | 0.00106 | 0.7555 |
| VAR(6) | 3220.158 | $7.542 \cdot 10^{-06}$ | 0.9955 |
| VAR(7) | 3201.728 | $5.378 \cdot 10^{-07}$ | 0.9035 |
| VAR(8) | 3137.745 | $3.575 \cdot 10^{-09}$ | 0.901 |
| VAR(9) | 3041.397 | $2.2 \cdot 10^{-16}$ | 0.8118 |
| VAR(10) | 2830.188 | $2.2 \cdot 10^{-16}$ | 0.8175 |

En la Tabla 3 se observan 3 pruebas realizadas a los modelos predictivos:

- BIC (Criterio de información Bayesiano): Este criterio determina la capacidad predictiva de cada uno de los modelos.
- Correlación Serial: Esta columna muestra el test de correlación serial que nos indica que las observaciones no están relacionadas entre sí.
- Normalidad JB: En esta columna se indican los valores del test Jarque Bera para los modelos, el nivel de significancia es 0.05 es decir que valores mayores indican que los errores cumplen el supuesto de normalidad es decir que se distribuyen normalmente.

Los modelos que tienen mejores parámetros son los VAR (6), VAR(7), VAR(8) para cada uno de estos modelos se realizó el diagnóstico de residuales de los cuales se observa que el modelo que se ajusta mejor a cada una de las series es el modelo VAR(8), para este modelo se muestran los resultados a continuación:

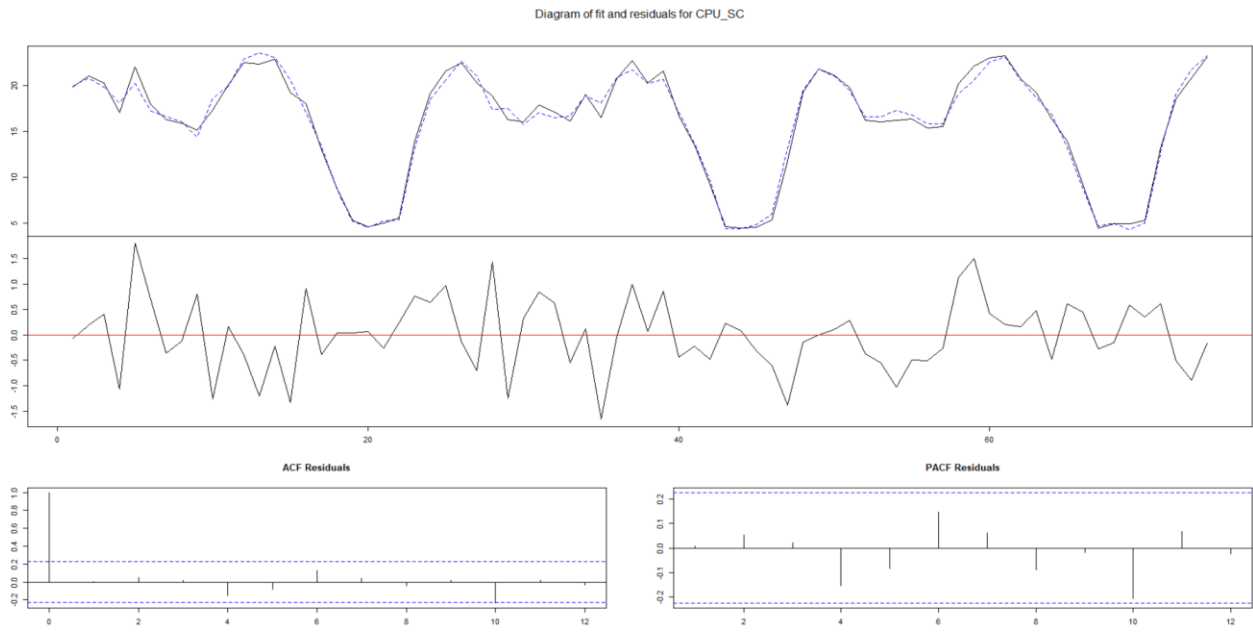


Ilustración 8. Residuales para la variable CPU_SC modelo VAR(8).

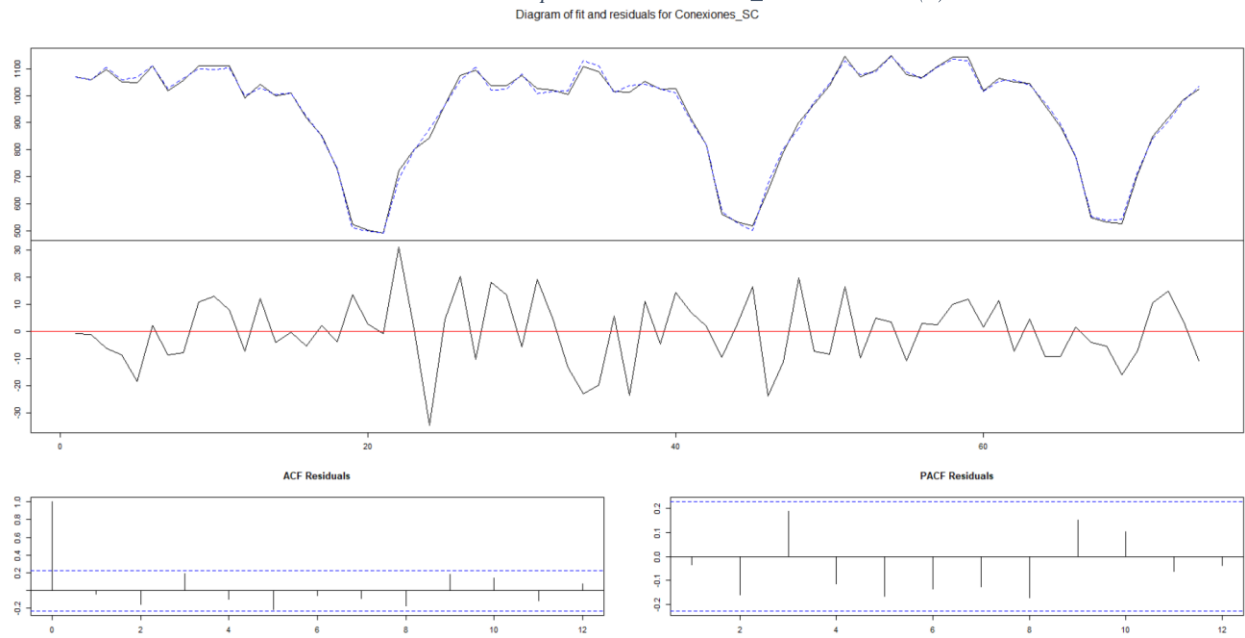


Ilustración 9. Residuales Conexiones SC modelo VAR(8).

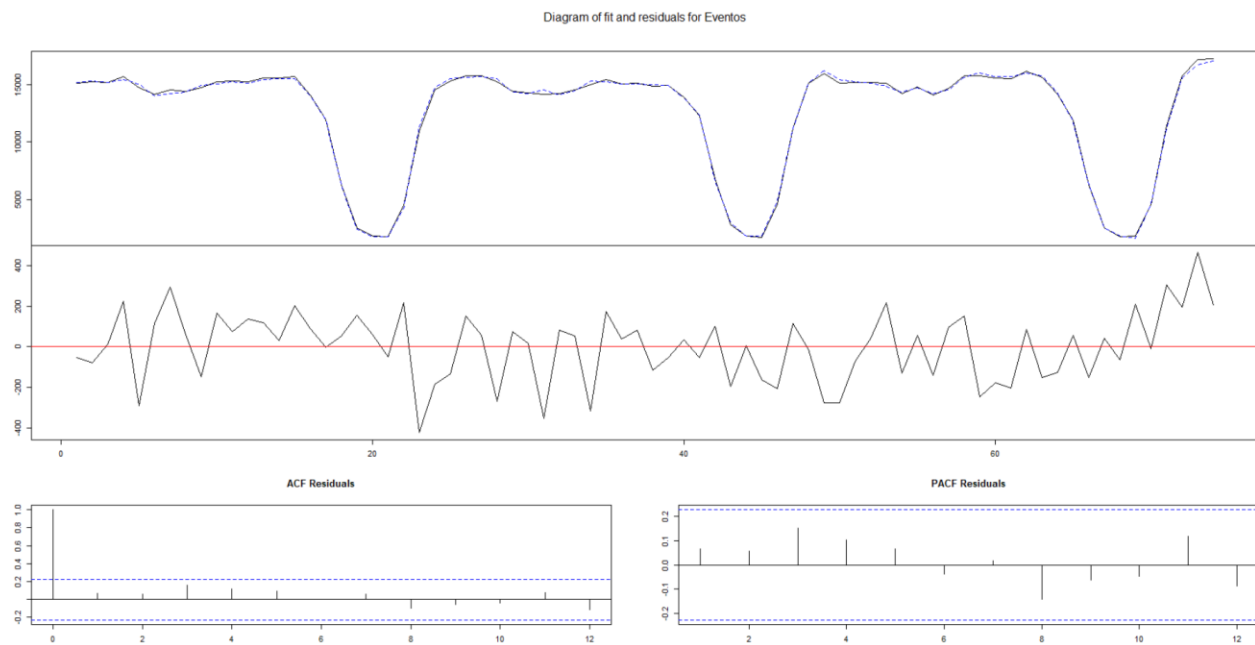


Ilustración 10. Residuales Eventos modelo VAR(8).

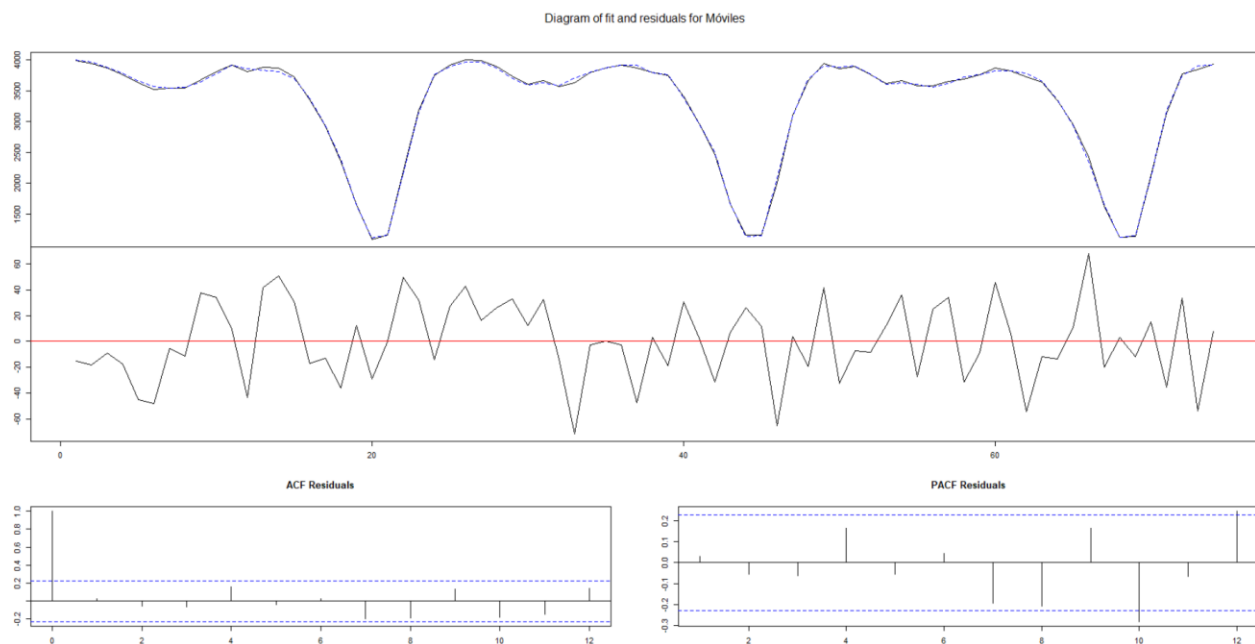


Ilustración 11. Residuales Móviles modelo VAR(8).

Luego se realizó el test de ljung box a los residuales y se realizó la gráfica qq plot para verificar que todos las observaciones se distribuyen de manera normal por el modelo VAR(8).

- CPU_SC (CPU del sistema central).

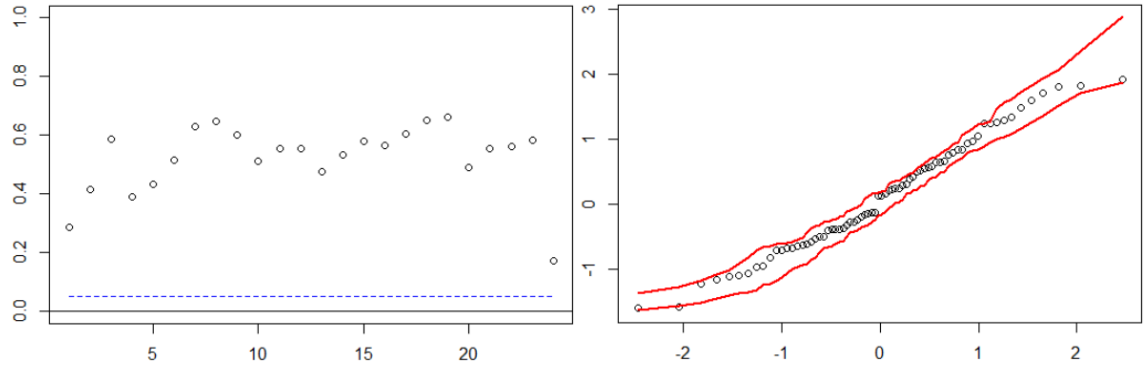


Ilustración 12. Diagnostico Ljung-Box y QQplot para los residuales de CPU_SC.

- Conexiones SC (Conexiones al sistema central).

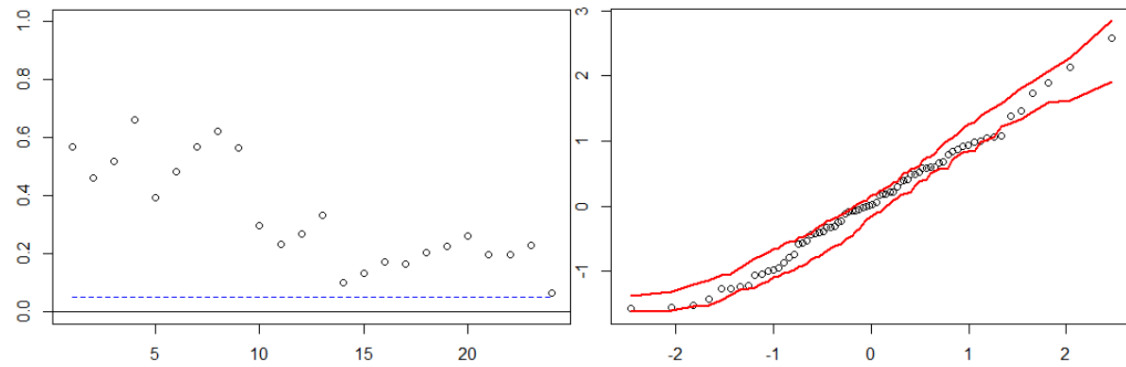


Ilustración 13. Diagnostico Ljung-Box y QQplot para los residuales de Conexiones SC.

- Eventos.

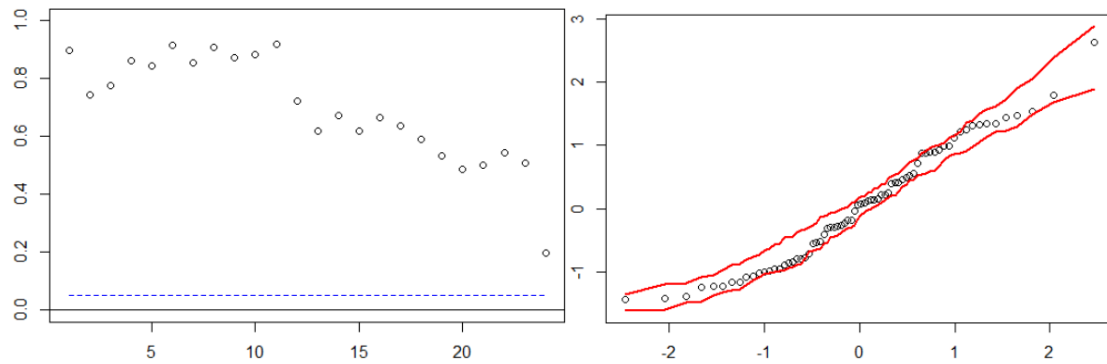


Ilustración 14. Diagnostico Ljung-Box y QQplot para los residuales de Eventos.

- Móviles.

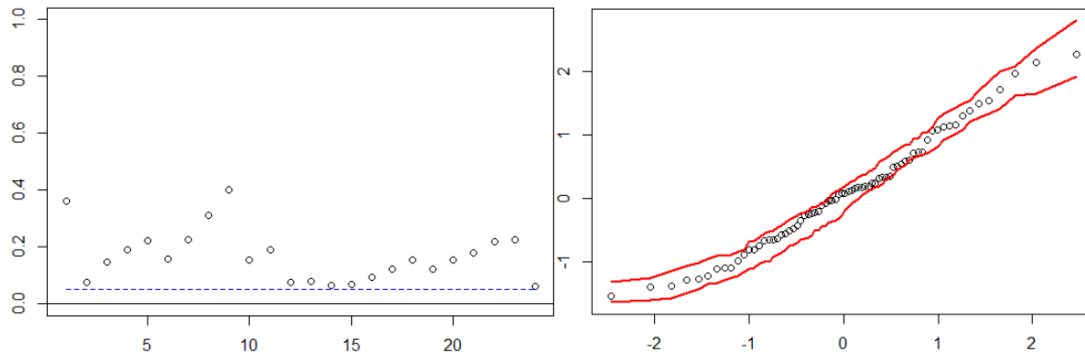


Ilustración 15. Diagnostico Ljung-Box y QQplot para los residuales de Móviles.

Se observan que los residuales cumplen con test de Ljung-Box, lo que nos permite afirmar que las observaciones son independientes, esto quiere decir que nuestro modelo predictivo tendrá una exactitud confiable y además en los diagramas de QQplot se concluyen que los residuales están dentro de las bandas de normalidad con un 5% precisión.

5.6. Funciones Impulso respuesta.

Graficar el comportamiento de las funciones impulso respuesta es importante en este trabajo ya que nos muestra los efectos transitorios ante un cambio en las innovaciones de una de las series, además visualizamos de mejor manera las relaciones entre las variables.

A continuación se presentan las funciones impulso respuesta para cada una de las series.

- Respuesta del comportamiento de la CPU del sistema central a una innovación de las series.

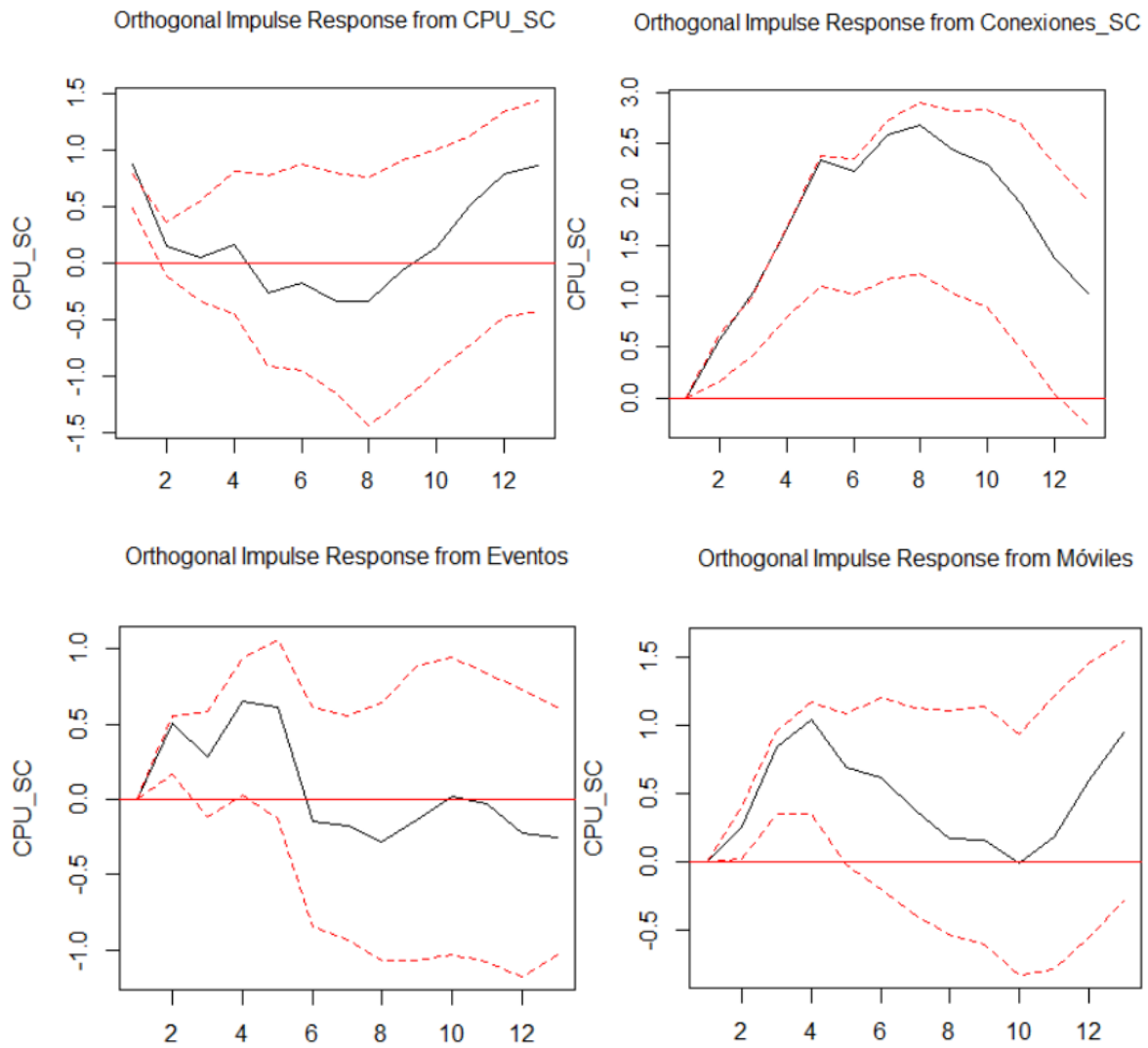
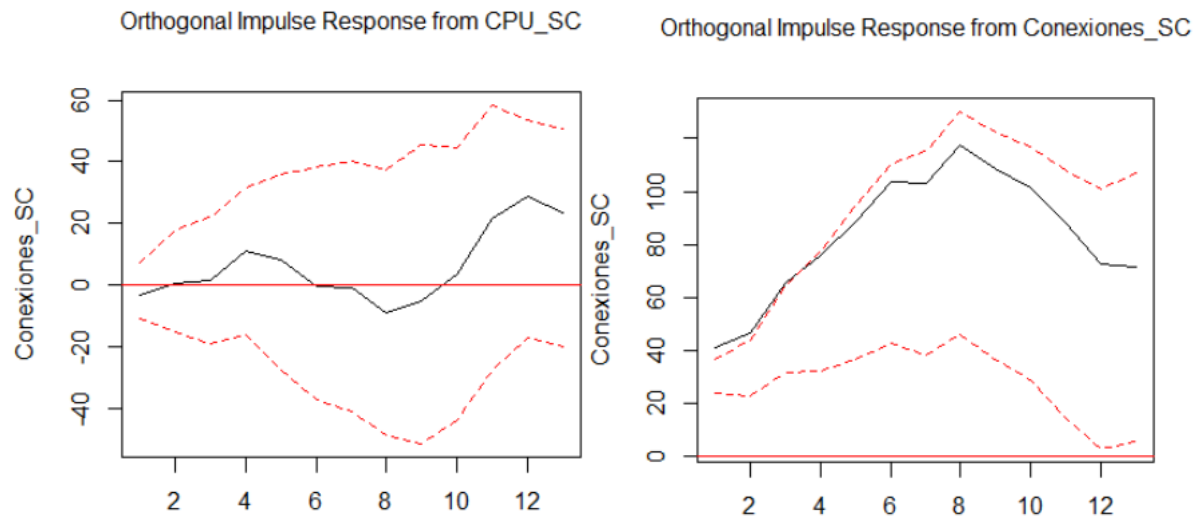


Ilustración 16. Respuesta que presentan la CPU del sistema central cuando se presenta innovación en otra serie.



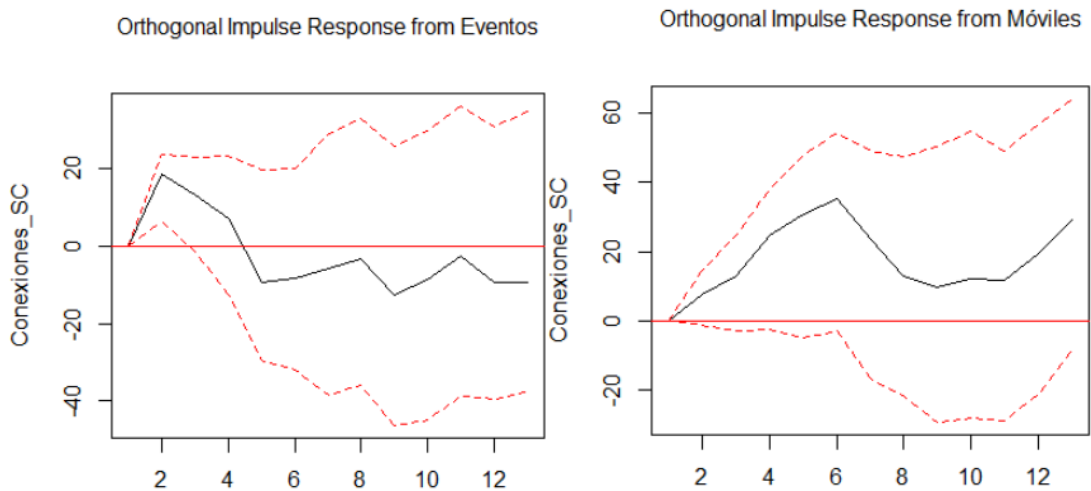


Ilustración 17. Respuesta que presentan a las conexiones cuando se presenta innovación en otra serie.

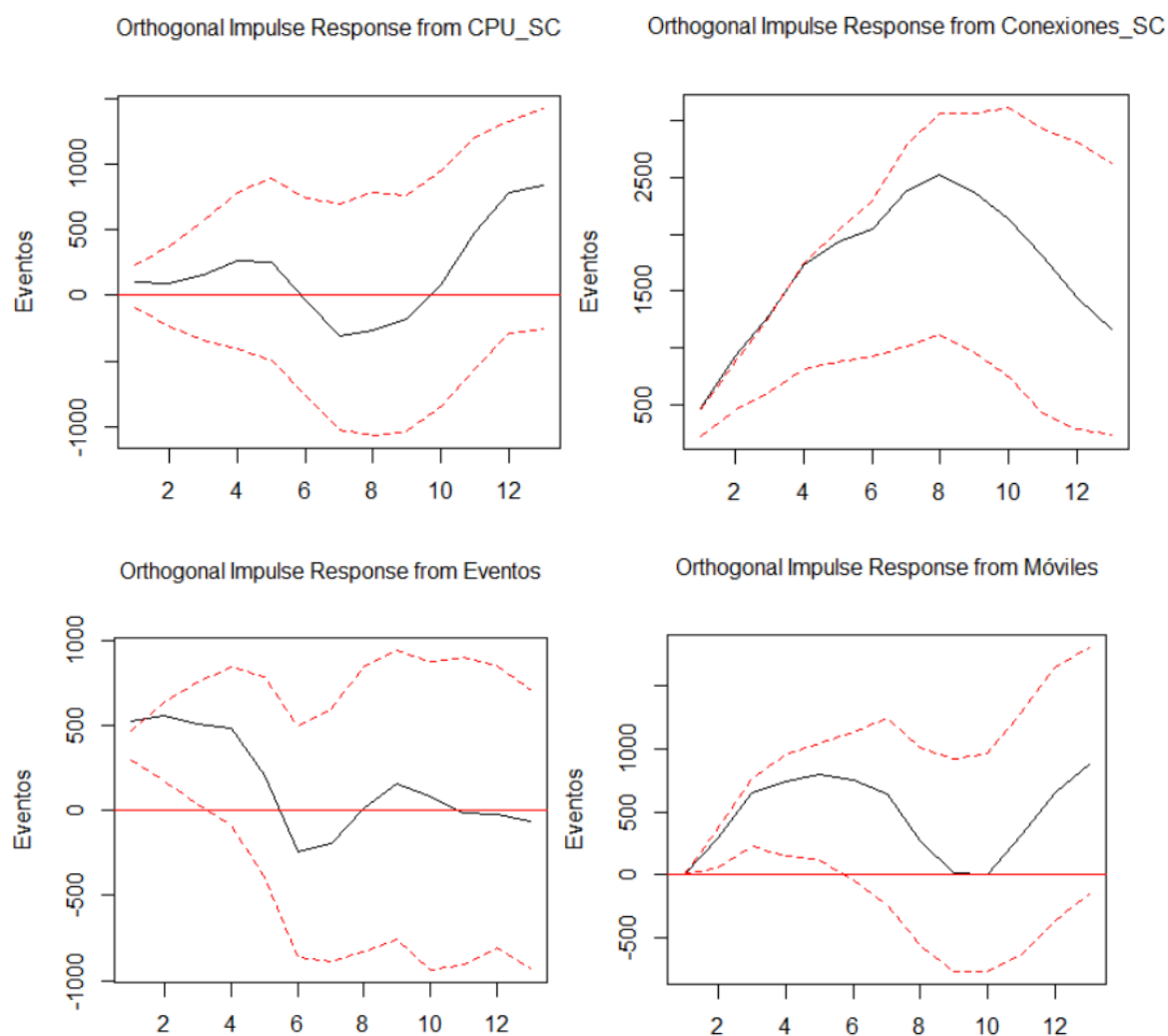


Ilustración 18. Respuesta que presentan los Eventos cuando se presenta innovación en otra serie.

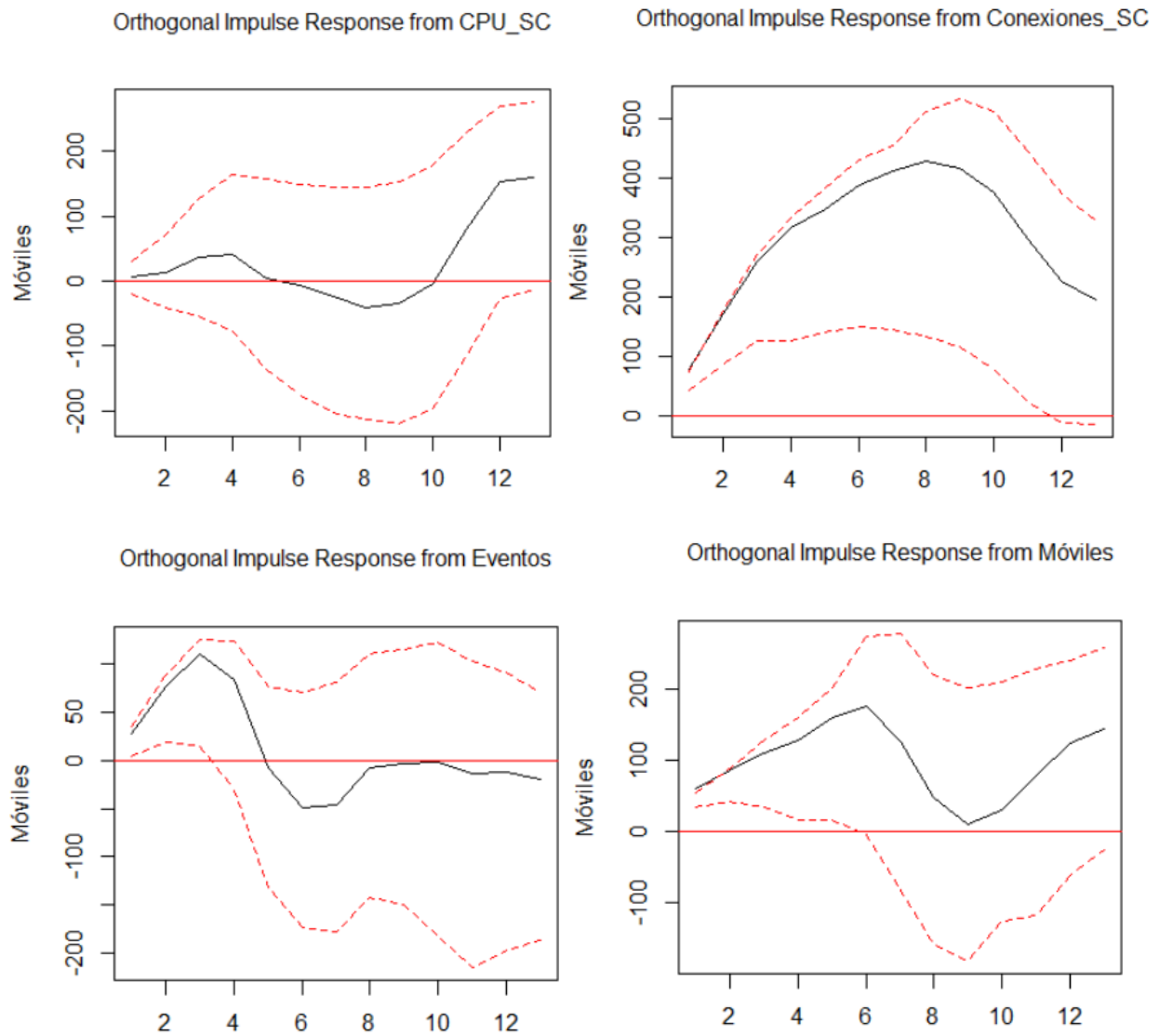


Ilustración 19. Respuesta que presentan los Móviles en el sistema cuando se presenta innovación en otra serie

5.7. Descomposición de varianza.

Al realizar la descomposición en varianza de las variables se puede observar de manera clara la incidencia que tiene cada variable sobre ella misma y la incidencia sobre las demás.

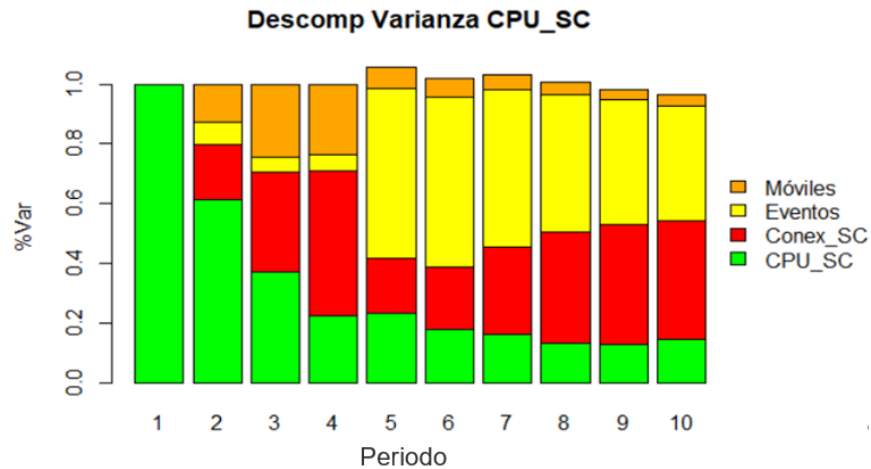


Ilustración 20. Descomposición de varianza para la CPU del sistema central..

Las conexiones al sistema central afectan de manera significativa la CPU del servidor del sistema central.

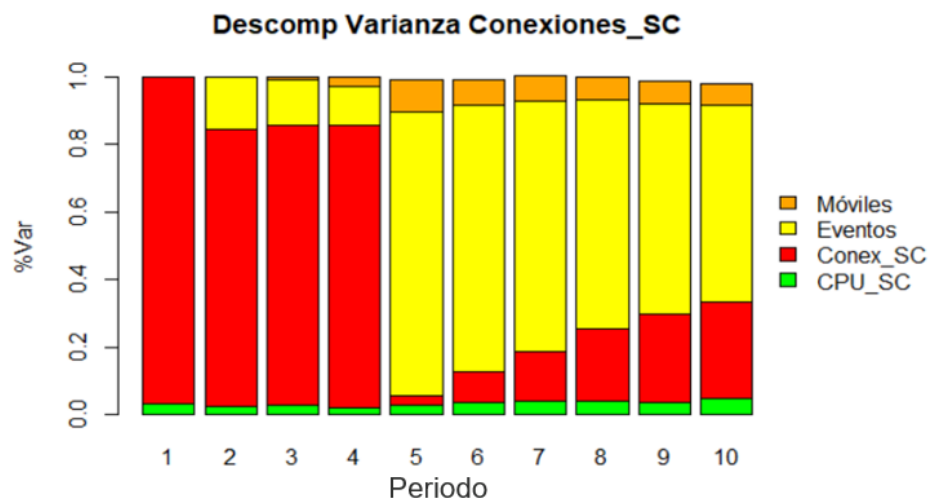


Ilustración 21. Descomposición de varianza para las Conexiones del sistema central..

El número de eventos presentados en el sistema incide de manera significativa las conexiones el en sistema central.

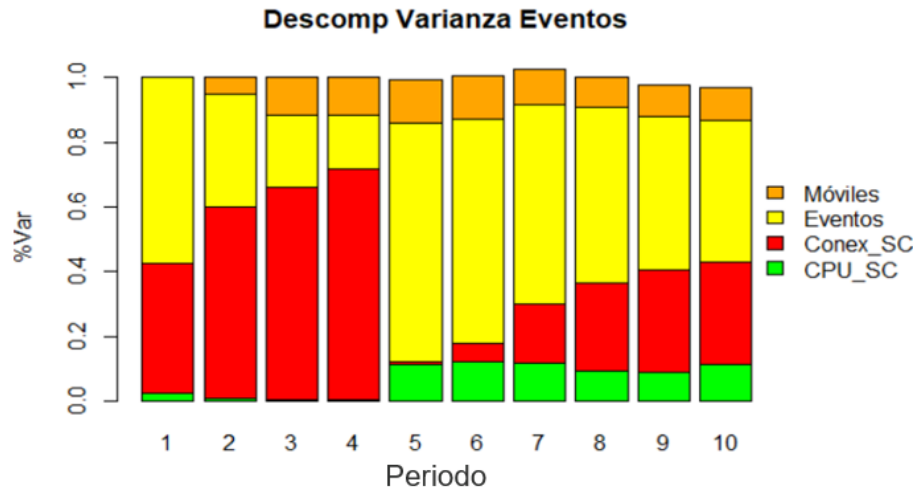


Ilustración 22. Descomposición de varianza para los Eventos del sistema.

La cantidad de eventos es afectada principalmente por el número de conexiones al sistema central.

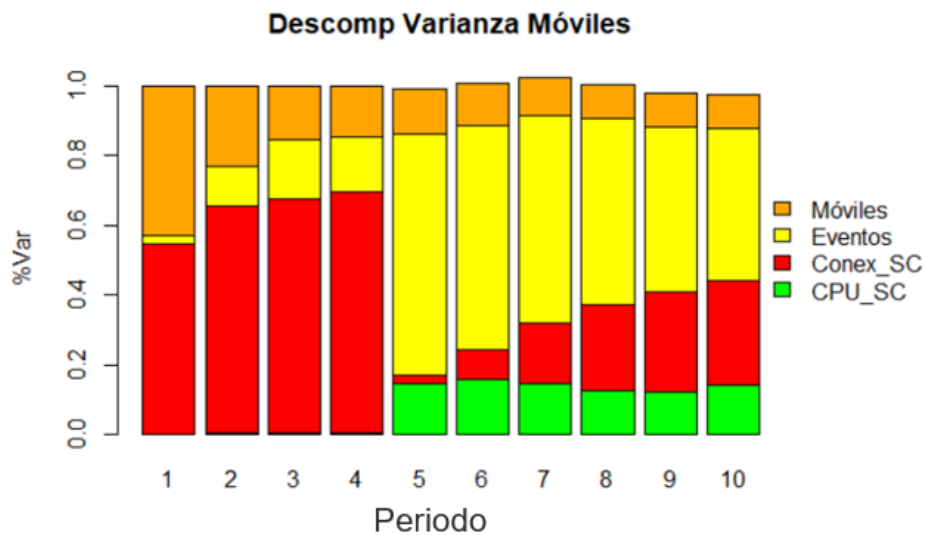


Ilustración 23. Descomposición de varianza para los Móviles del sistema.

La cantidad de móviles en operación es afectada de manera importante por el número de conexiones realizadas al sistema central.

De la descomposición de varianza se concluye que las variables a las que se debe monitorear de manera constante y prioritaria es el número de conexiones al sistema y la cantidad de eventos registrados ya que afectan de manera significativa a las demás variables propuestas y a sí mismas.

5.8. Ajuste del modelo propuesto.

- Outsample test:

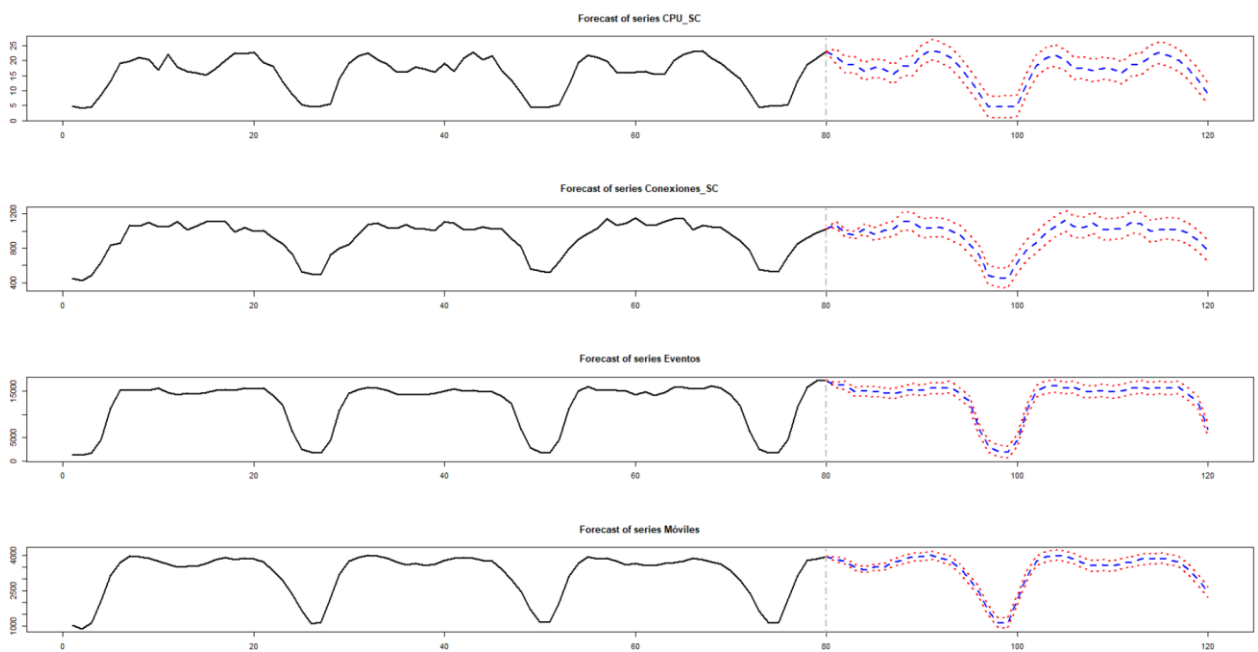
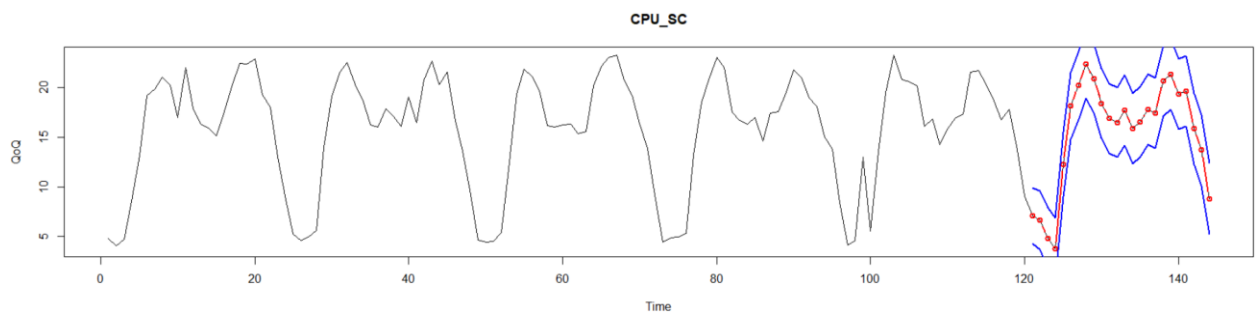


Ilustración 24. Test Outsample para el modelo VAR(8).

5.9. Ajuste del modelo propuesto.



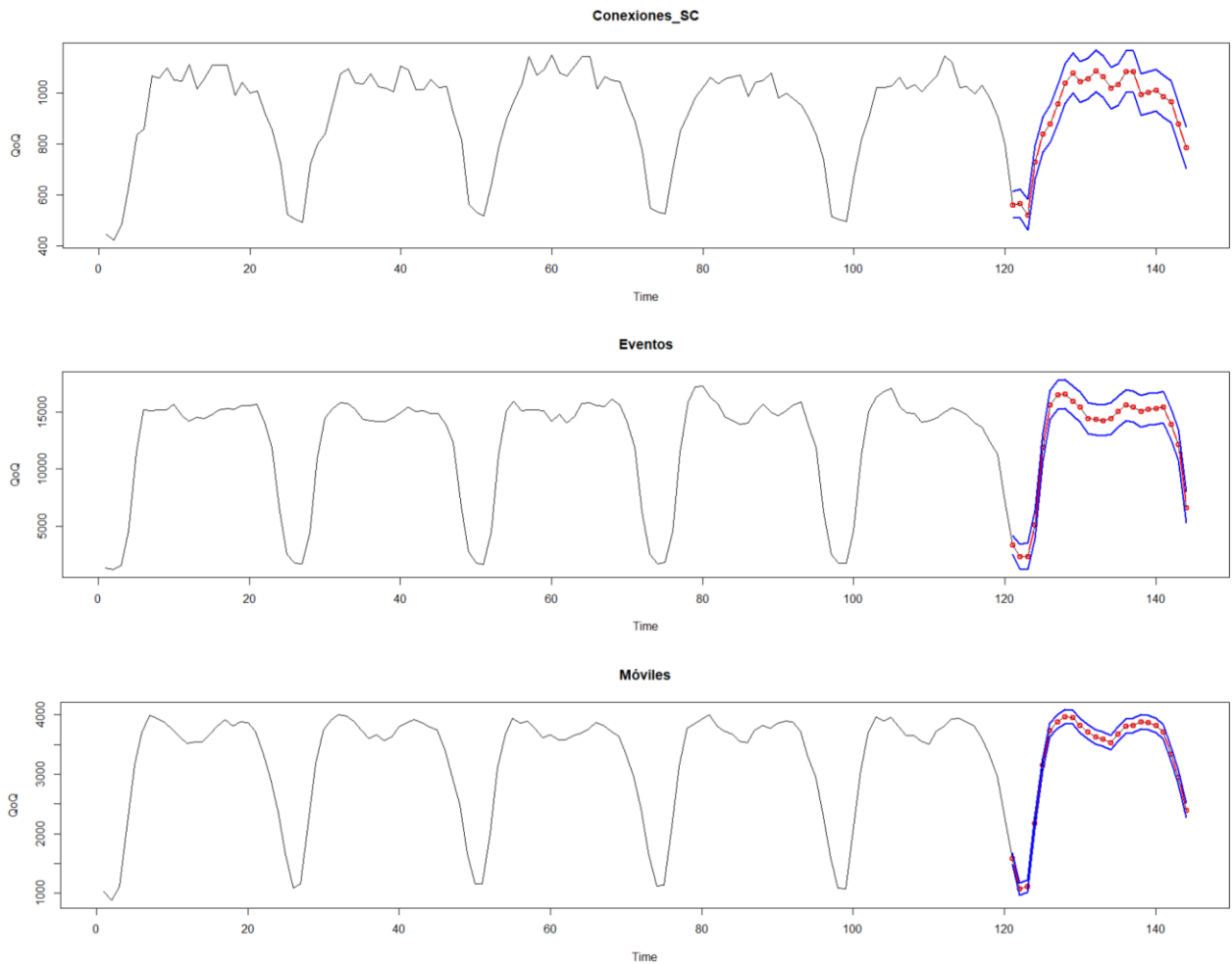


Ilustración 25. Test outsample para cada una de las variables.

Se observa que estos valores son razonables en la práctica y se ajustan al comportamiento del sistema.

5.10. Ecuaciones de cada una de las variables del modelo propuesto.

Endogenous variables: CPU_SC, Conexiones_SC, Eventos, Móviles

Deterministic variables: const

Sample size: 74

Log Likelihood: -1476.649

Roots of the characteristic polynomial:

0.9506 0.9506 0.942 0.942 0.8832 0.8832 0.8737 0.8737 0.8525 0.8525 0.8496 0.8496 0.8442
0.8442 0.8415 0.8415 0.819 0.819 0.7664 0.7664 0.7502 0.7502 0.525 0.525

Call:

vars::VAR(y = s2_New, p = 6)

Equation CPU_SC:

$$CPU_SC = CPU_SC.l1 + Móviles.l1 + Eventos.l2 + CPU_SC.l4 + Eventos.l6 + const$$

| Estimation results for equation CPU_SC | | | | | |
|--|-----------|------------|---------|----------|-----|
| Coef | Estimate | Std. Error | t value | Pr(> t) | |
| CPU_SC.l1 | 3.12E-01 | 1.38E-01 | 2.262 | 0.028149 | * |
| Móviles.l1 | 7.78E-03 | 2.12E-03 | 3.678 | 0.000585 | *** |
| Eventos.l2 | -1.14E-03 | 3.41E-04 | -3.359 | 0.001521 | ** |
| CPU_SC.l4 | -4.05E-01 | 1.41E-01 | -2.876 | 0.005946 | ** |
| Eventos.l6 | 7.41E-04 | 3.27E-04 | 2.269 | 0.027705 | * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.329 on 49 degrees of freedom
Multiple R-Squared: 0.966, Adjusted R-squared: 0.9493
F-statistic: 57.94 on 24 and 49 DF, p-value: < 2.2e-16

Equation Conexiones_SC:

$$Conexiones_SC = Eventos.l1 + CPU_SC.l5 + Conexiones_SC.l6 + Eventos.l6 + const$$

| Estimation results for equation Conexiones_SC | | | | | |
|---|----------|------------|---------|----------|-----|
| | Estimate | Std. Error | t value | Pr(> t) | |
| Eventos.l1 | 0.041016 | 0.010704 | 3.832 | 0.000363 | *** |
| CPU_SC.l5 | -17.8784 | 5.876123 | -3.043 | 0.003764 | ** |
| Conexiones_SC.l6 | - | | | | |
| | 0.459578 | 0.14306 | -3.212 | 0.002327 | ** |
| Eventos.l6 | 0.02983 | 0.011932 | 2.5 | 0.015812 | * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.57 on 49 degrees of freedom
Multiple R-Squared: 0.9562, Adjusted R-squared: 0.9347
F-statistic: 44.52 on 24 and 49 DF, p-value: < 2.2e-16

Equation Eventos:

$$Eventos = Eventos.l1 + Móviles.l1 + Eventos.l2 + const$$

| Estimation results for equation Eventos | | | | | |
|---|----------|------------|---------|----------|-----|
| | Estimate | Std. Error | t value | Pr(> t) | |
| Eventos.l1 | 0.7668 | 0.1964 | 3.904 | 0.00029 | *** |
| Móviles.l1 | 4.1832 | 1.4194 | 2.947 | 0.004899 | ** |
| Eventos.l2 | -0.8002 | 0.2284 | -3.504 | 0.000989 | *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 891.4 on 49 degrees of freedom
Multiple R-Squared: 0.9768, Adjusted R-squared: 0.9654
F-statistic: 85.83 on 24 and 49 DF, p-value: < 2.2e-16

Equation Móviles:

$Móviles = Eventos.l1 + Móviles.l1 + CPU_SC.l4 + Móviles.l6 + const$

| Estimation results for equation Móviles | | | | | |
|---|-----------|------------|---------|----------|-----|
| | Estimate | Std. Error | t value | Pr(> t) | |
| Eventos.l1 | 0.09469 | 0.02847 | 3.326 | 0.00168 | ** |
| Móviles.l1 | 1.07737 | 0.2057 | 5.238 | 3.41E-06 | *** |
| CPU_SC.l4 | -43.87773 | 13.68703 | -3.206 | 0.00237 | ** |
| Móviles.l6 | -0.51248 | 0.20524 | -2.497 | 0.01594 | * |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.2 on 49 degrees of freedom
Multiple R-Squared: 0.9852, Adjusted R-squared: 0.9779
F-statistic: 135.6 on 24 and 49 DF, p-value: < 2.2e-16

Capítulo 6 Conclusiones y recomendaciones

6.1. Conclusiones

De acuerdo al análisis de resultados obtenido se concluyó lo siguiente:

- Se obtuvieron 4 variables generales que describen el comportamiento general de un sistema de transporte.
- El modelo propuesto VAR (8) describe de manera acertada las 4 variables usadas para el desarrollo del trabajo.
- Se encontraron que las variables que más son susceptibles a las demás y sobre las cuales debemos prestar más atención en nuestro análisis son “Número de Conexiones” y “Cantidad de Eventos” ya que estas inciden de manera importante a cada una de las demás variables.
- Se realizó el pronóstico para 24 muestras para el comportamiento de cada una de las variables.

6.2. Recomendaciones

- Evaluar la mayor cantidad de variables de las cuales se cuente con información detallada e histórica, esto con el fin de conseguir un modelo perfectamente ajustado y con poco margen de error de predicción.
- Integrar una aplicación web para el análisis en tiempo real y envío de alertas a las personas encargadas de soporte y monitoreo.
- Realizar un script que permita ejecutar este procedimiento de manera regular para usarla como herramienta de monitoreo.

- Analizar con modelos VARMA y verificar ajuste y pronóstico de cada una de las variables.

Capítulo 7 Referencias bibliográficas

7.1. Referencias

Sowell, B., (2007). Big Data Para la Movilidad. <http://www.aacarreteras.org.ar>

Ciudades en movimiento. (2016). Revisión de la estrategia del transporte urbano del banco mundial.

Correa Santiago (2012). Transporte y desarrollo urbano en Colombia: los tranvías de Bogotá y Medellín. <http://editorialcesa.com/media/preview/transporte-y-desarrollo-economico-en-colombia.pdf>.

CAF (2017). Desarrollo urbano y America Latina. https://www.caf.com/media/4203/desarrollourbano_y_movilidad_americalatina.pdf

Novales A, (2014). Modelos Vectoriales Autoregresivos (VAR). Universidad Complutense.

Londoño W, (2005). Modelos de ecuaciones multiples, modelos AVR y cointegración. Universidad EAFIT.